STAT306: Group Project

Investigating the Most Influential Factors on House Prices

Instructor: Xinglong Li

Group Members (B7): Lily Xie, Maxwell Woodfield, Maisie Wu, Yanjun Wang

Introduction

In recent decades, the real estate industry has exhibited a sustained and continuous expansion. Ever since much of the industrialized world experienced the unprecedented rise in house prices in the late 1990s and early 2000s (Kim K., 2008), the upward trajectory in residential property values has remained unrelenting, especially in metropolitan cities. Our incentive for analyzing this dataset is to disentangle the enigma surrounding the elevated real estate prices. Specifically, we want to investigate the associations between each of the explanatory variables and house price and to determine which of these factors influences house price the most. Furthermore, we will also utilize these data to construct an inferential model capable of explaining the relationship between house price and various explanatory variables in the dataset, based on the available information.

The data selected in this model is a Real Estate valuation data set, which can be found here: Real Estate Valuation Data Set. This market historical dataset of real estate valuation was collected in 2013 from Sindian Dist., New Taipei City, Taiwan, and was donated to the UC Irvine Machine Learning Repository in 2018. The dataset consists of 414 observations including 6 explanatory variables, and 1 response variable outlined below. In this model, we aim to draw conclusions about the most critical variables, as well as how they interact in influencing house prices.

Variable	Unit	Mean	S.D	Median	Min	Max
Transaction date	N/A (Note: 2013.250=2013 March, 2013.500=2013 June, etc)	N/A	N/A	N/A	N/A	N/A
House age	Year	17.71	11.39	16.10	0.00	43.80

Distance to the nearest MRT(Metro system) station	Meter	1083.89	1262.11	492.23	23.38	6488.02
Number of convenience stores within walking distance	Stores	4.09	2.95	4.00	0.00	10.00
Geographic coordinates (latitude)	Degree (North/South)	24.97	0.01	24.97	24.93	25.01
Geographic coordinates (longitude)	Degree (West/East)	121.53	0.02	121.54	121.47	121.57
House price per unit area	10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared	37.98	13.61	38.45	7.60	117.50

Table 1: Variables in Real Estate Valuation Data Set

Analysis

1. Exploratory Data Analysis and Key Features

First, we would like to read in and obtain a brief overview of the data.
After reading the data, we decided to give the column names more descriptive titles and remove an unnecessary column. You can refer to this data-cleaning process in the appended R code file.

It is also a good practice to check and remove any missing values in the data. Fortunately, no missing observations have been found in this dataset.

And we can start performing our preliminary data analysis now.

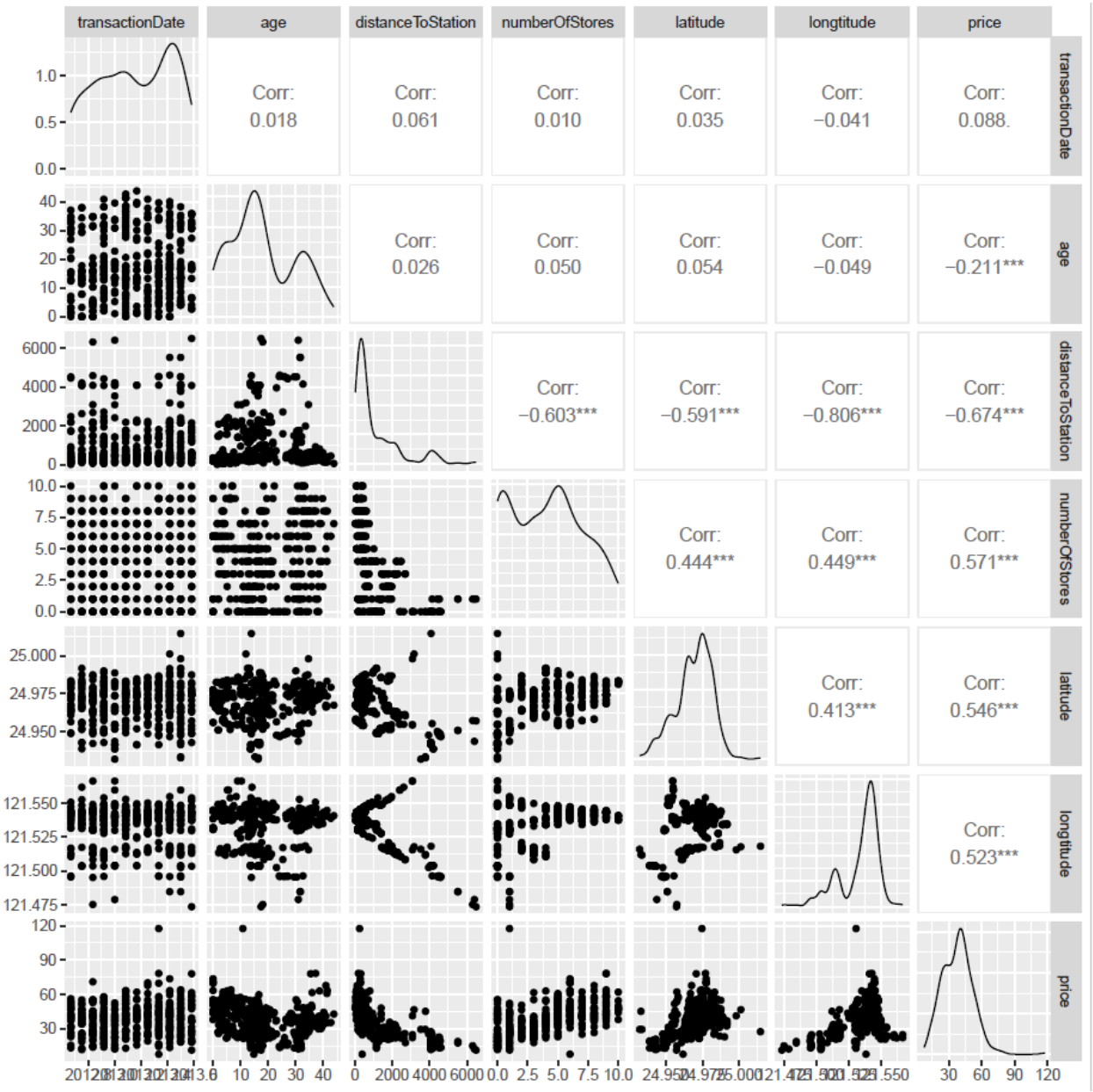

Figure 1: A matrix of plots showing pairwise relationships between raw variables in the data set

We start by looking at the relationships between all the variables we are interested in from the raw dataset using ggpairs() function. According to the generated plot as shown in Figure 1, the distance to the nearest MTR station exhibits a notable negative correlation with the

price, reflected by a correlation coefficient of -0.674. Additionally, the number of stores nearby, latitude and longitude all demonstrate moderately strong positive linear associations with the price, with correlation coefficients of 0.571, 0.546 and 0.523, respectively.

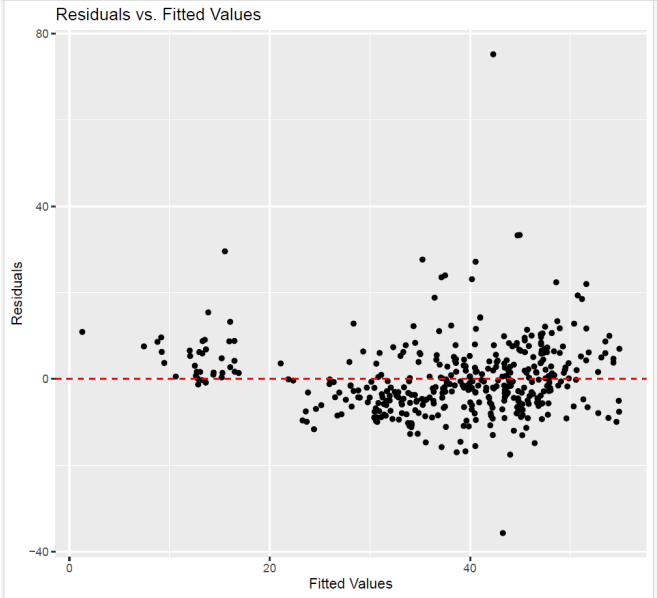

Figure 2: Scatterplot with residuals over fitted values from the model involving all predictors

By constructing a scatter plot with the x-axis depicting the predicted (fitted) values derived from the complete regression model (using all the variables other than price as the linear predictors in the raw data, without interaction terms), and the y-axis illustrating the residuals, we look for the residuals for discernible patterns or trends (figure 2). Ideally, the data points should exhibit a random dispersion around the horizontal line at zero. However, in the presented figure, the points do not display a random distribution around the horizontal zero line, suggesting the possibility of patterns within the residuals, and potential violation of the linear assumption if we were to fit a model without manipulations on the data.

2. Analyzing Potential Predictors

We have noticed that the transaction date in this dataset is encoded in a way that the year and month belong to the same column, separated by decimal points and the month is encoded

specially. We therefore would convert the transaction date into year and month with separate columns.

After further cleaning the data, we would like to then run ggpairs() again, to find additional insight with year and month wrangled.

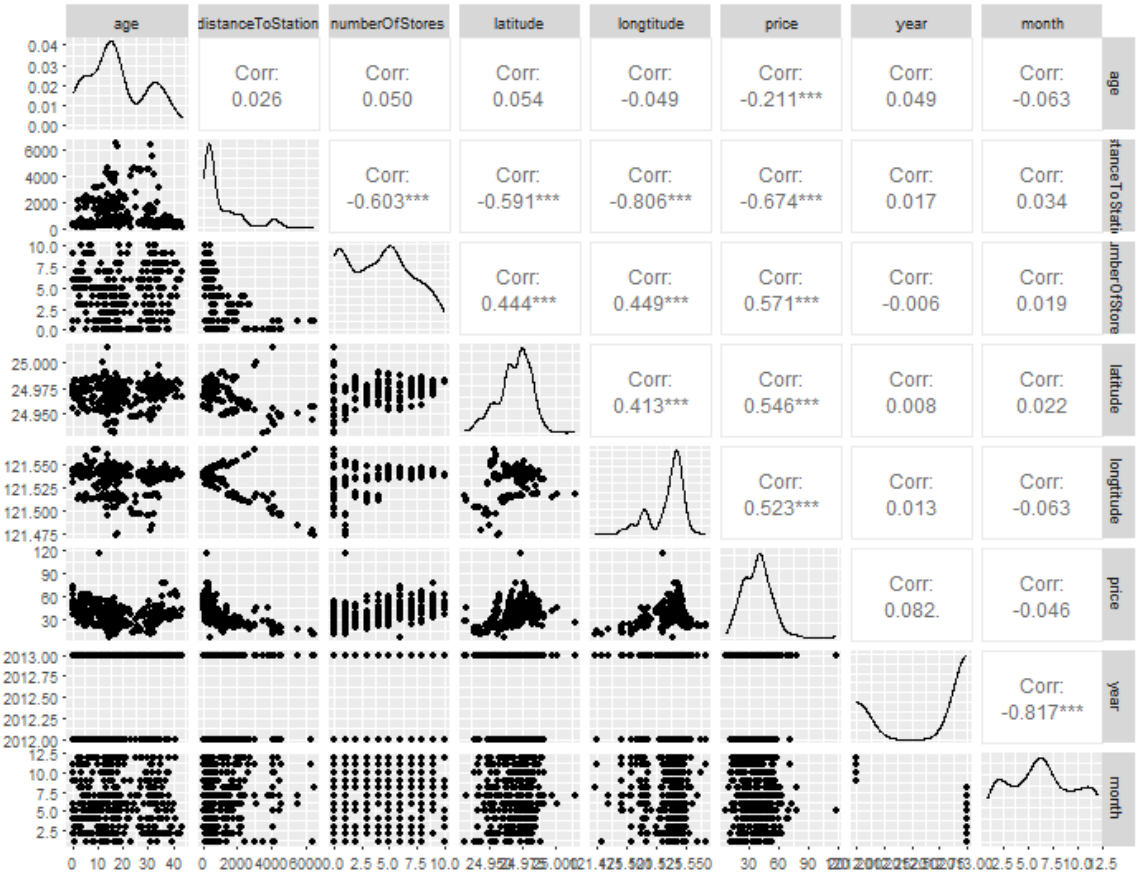

Figure 3: A matrix of plots showing pairwise relationships between cleaned variables in the data set

Based on the plot, price is significantly correlated with every other variable, so we would keep all the variables as predictors as a starting point.

It is worth noting that other than the relationships between price and other predictors, there is a strong and significant negative correlation between longitude and distance to the nearest MRT station (r=-0.806), with the highest absolute correlation coefficient value among all pairs (following that of year and month, which does not make much sense to include). Therefore, we would like to consider the interaction between the longitude and distance to the station in our later analysis. It was shown in the article that analysts use longitude to calculate the distance (Le, 2021).

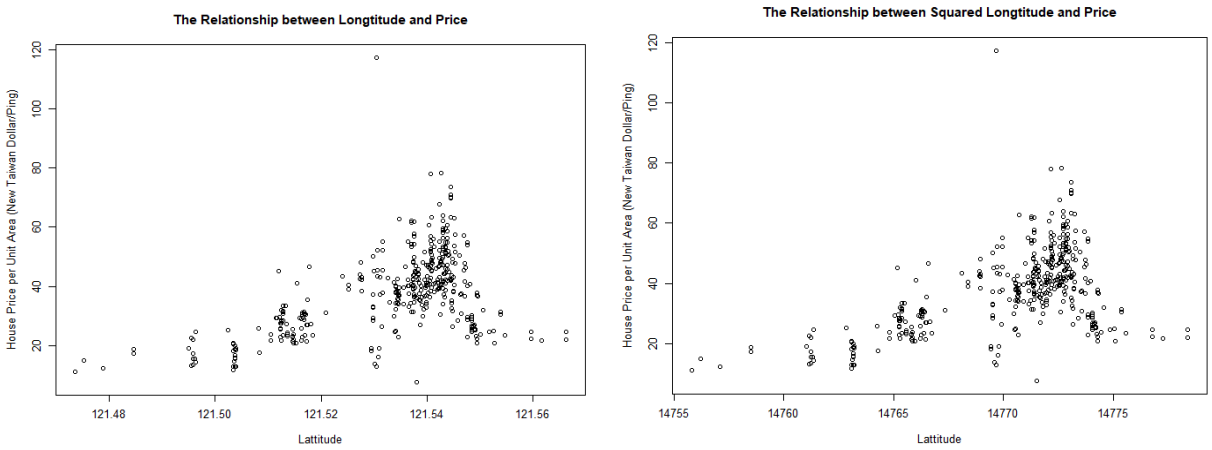

Figure 4: Relationship between Longitude/Squared Longitude and Price

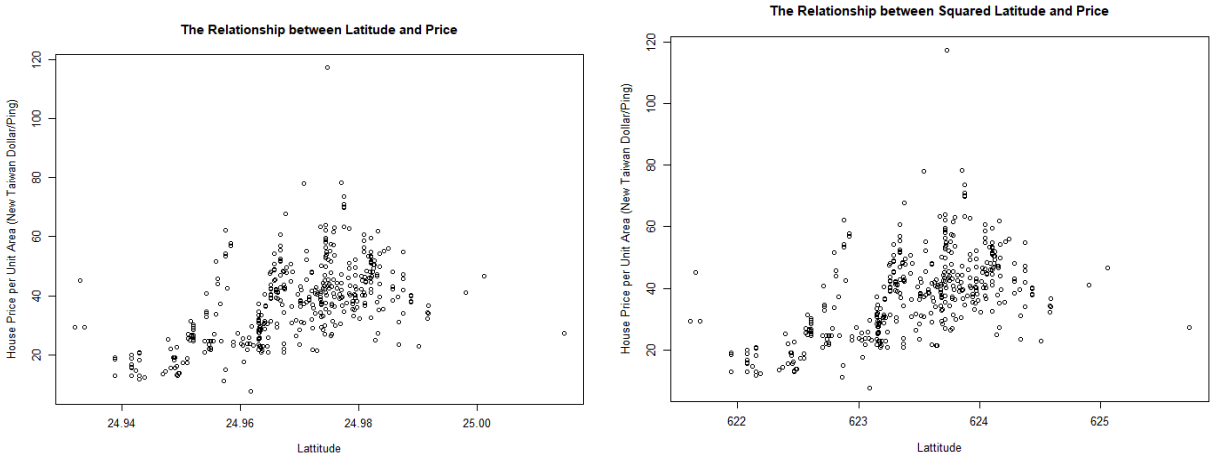

Figure 5: Relationship between Latitude/Squared Latitude and Price

By looking at the relationship between latitude and price as well as longitude and price, they seem to be correlated but not linearly: non-constancy of variance seems to violate our assumption of homoscedasticity under the linear assumption. Therefore it is worth considering whether to scale the predictors or the dependent variable.

By taking the squares of longitude and latitude, no significant change was found based on the plot in terms of fitting; scaling the price by taking the natural log of it would then make the price's originally strongly correlated relationship with distance to station no longer significant. Thus we would continue with these original variables.

By looking at the mean unit house price of 2012 and 2013 (36.3 and 38.7, respectively), the mean price of 2013 was significantly higher. Since there are only 2-year values possible, we convert the year into categorical values.

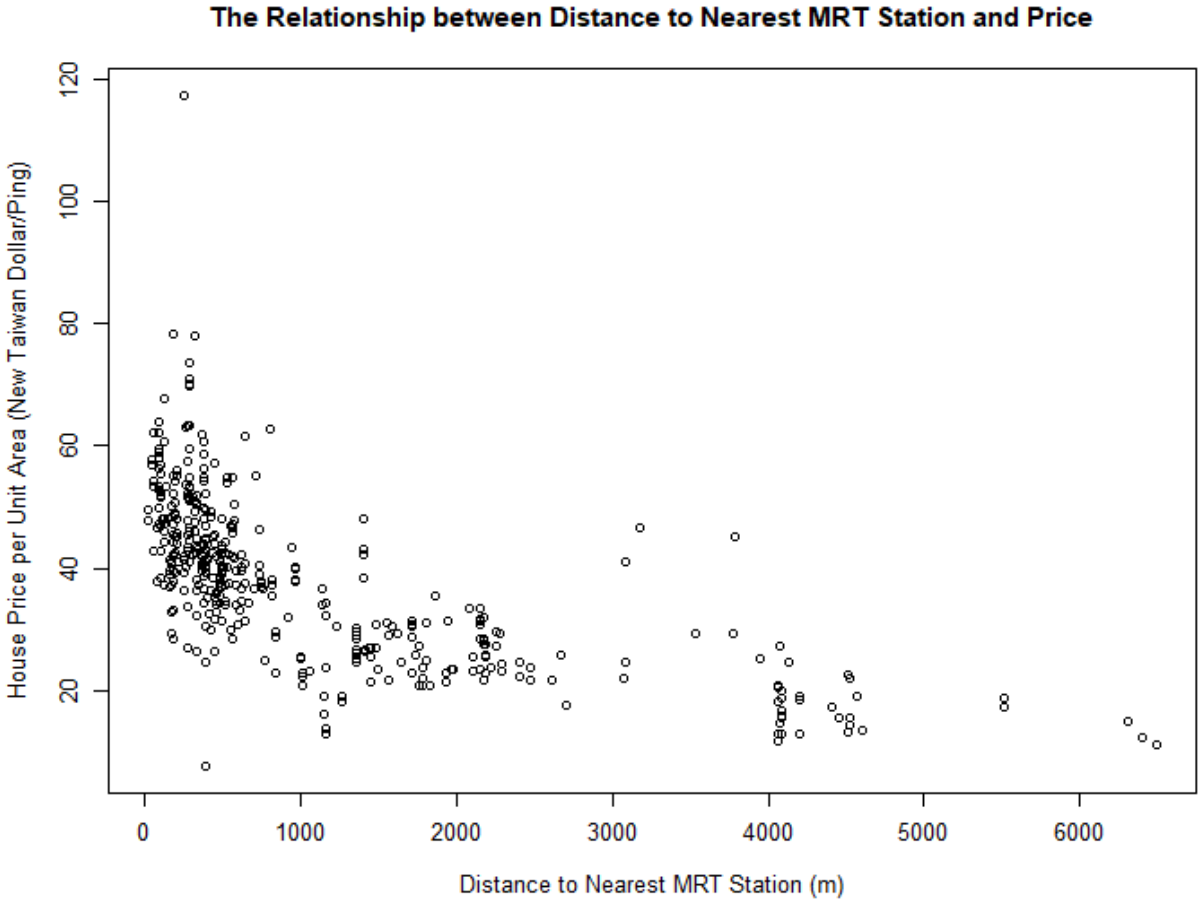

Figure 6: Relationship between Distance to Nearest MRT Station (m) and Price

Again, by plotting the unit house price against distance to the nearest MRT station, we have found a negative exponential relationship between distance and price. We would like to proceed as current since in the previous step we have experienced that scaling the dependent variable might affect its significance correlation with other predictors drastically.

3. Model Fitting and Diagnostics

With the wrangled variables in the dataset, we are now ready to fit linear models to explore the factor(s) that explain or influence the unit house price the most.

We will fit an all-inclusive model based on our preliminary analysis, then adjust the fitting using different variables according to the result of the first model, and finally, use Mallows's Cp statistics to help us navigate the best model, and compare these models together.

Based on the significant correlation coefficients between all of the variables in our preliminary data analysis, we would like to include all of our predictors in the first model we attempt to fit. We would also like to include the interaction term between longitude and distance to the nearest MRT station, as we have found them to have the strongest correlation among all predictors.

Model 1, the full model with all the predictors and with an interaction term between distance to station and longitude.

Model 1 involves all the variables (listed above) and the interaction term between distance to station and longitude, all the variables are significant except for the month (i.e. transaction date). The adjusted R square is 0.6099, which means the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables is 60.99%. The coefficient of residual standard error is 8.499. As the month variable is insignificant, we proceed to the second model without the month as an independent variable.

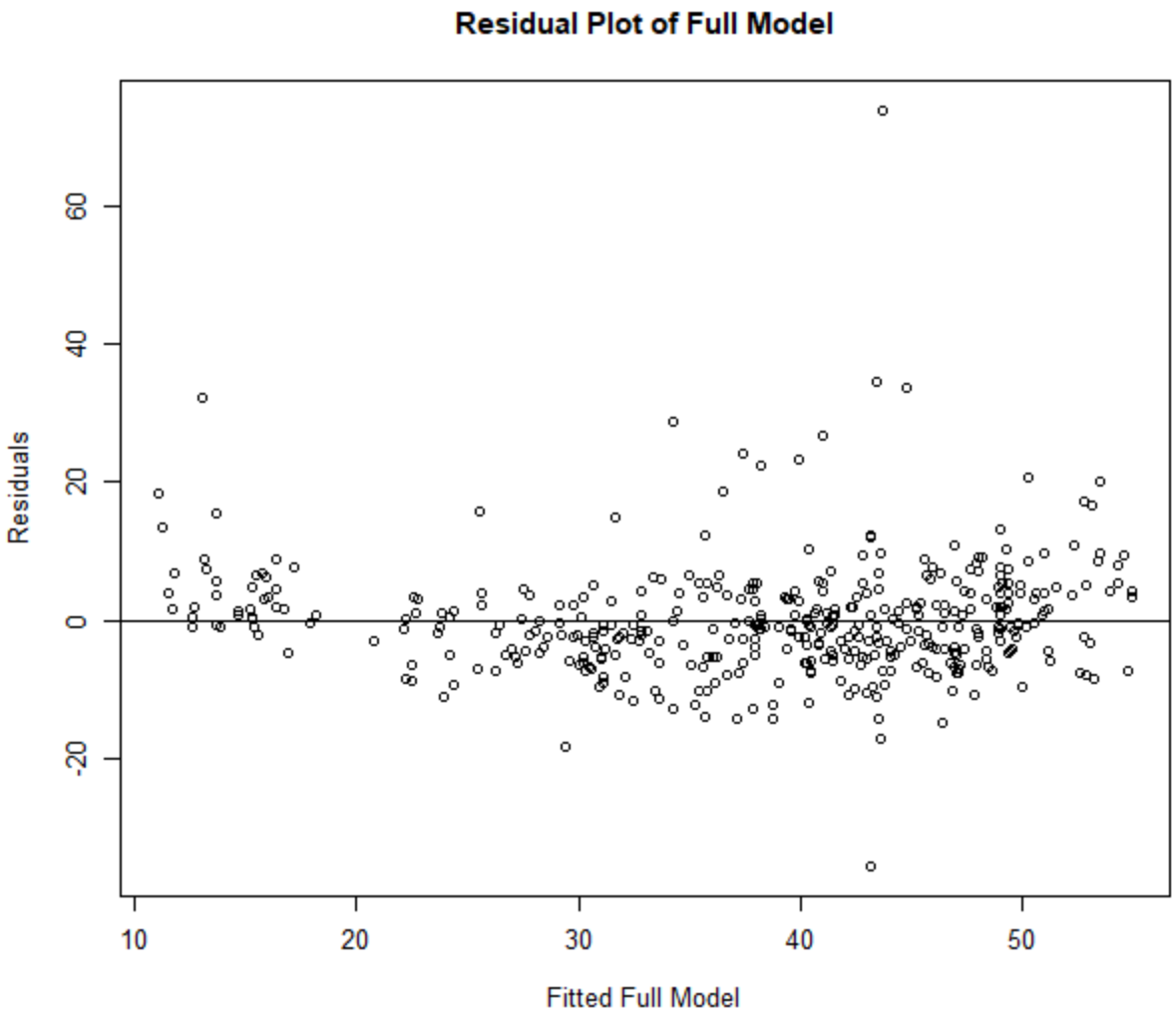

Figure 7: Scatterplot with residuals over fitted values from Model 1

Model 2, continues with the full model setup with all the predictors but removing month, and with an interaction term between distance to station and longitude

Model 2 leaves out the variable month. The adjusted R square is 0.6083, which is similar but slightly lower than model 1. However, the residual error is larger in model 2 compared to model

1, which is 8.516 > 8.499. Models fitted with smaller residual standard errors are preferred.

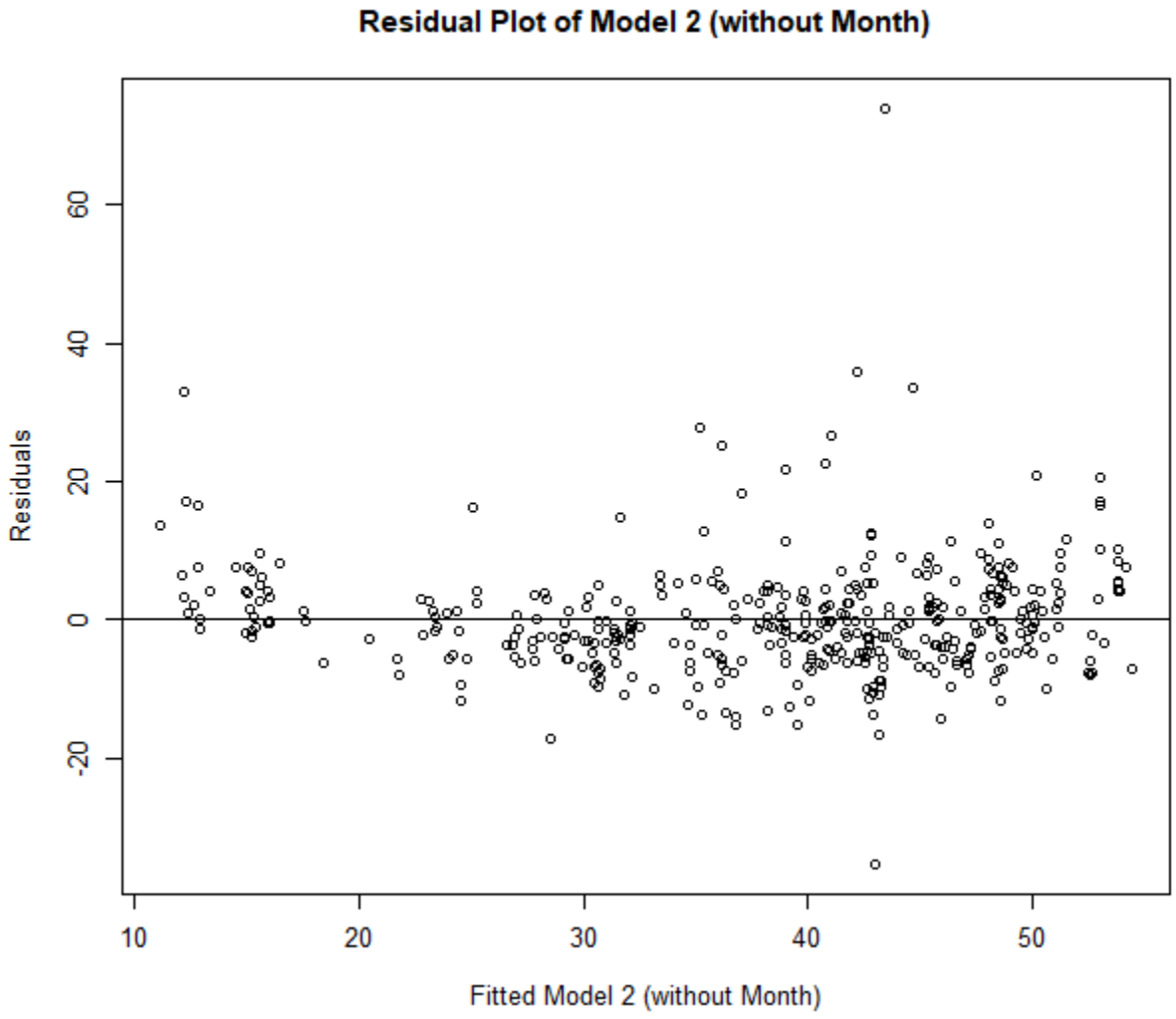

Figure 8: Scatterplot with residuals over fitted values from Model 2

These summary statistics for the two models are quite close to each other, so it would be difficult to tell using only the model summaries. Instead, we move on to step 3 and use Mallows's Cp statistics to assist us in determining the best model.

Model 3, Using Mallows's Cp to determine the best model

Below are the results of the best subsets regression using Mallows's Cp statistics. It includes the number of predictors present, and which variables should be used for the corresponding model. Note: "Interaction" stands for interaction between longitude and distance to the station below.

1: intercept
2: intercept, interaction
3: intercept, interaction, number of stores
4: intercept, interaction, age, distance to station
5: intercept, interaction, age, distance to station, latitude
6: intercept, interaction, age, distance to station, latitude, longitude
7: intercept, interaction, age, distance to station, latitude, longitude, number of stores
8: intercept, interaction, age, distance to station, latitude, longitude, number of stores, year
9: intercept, interaction, age, distance to station, latitude, longitude, number of stores, year, month

Cp

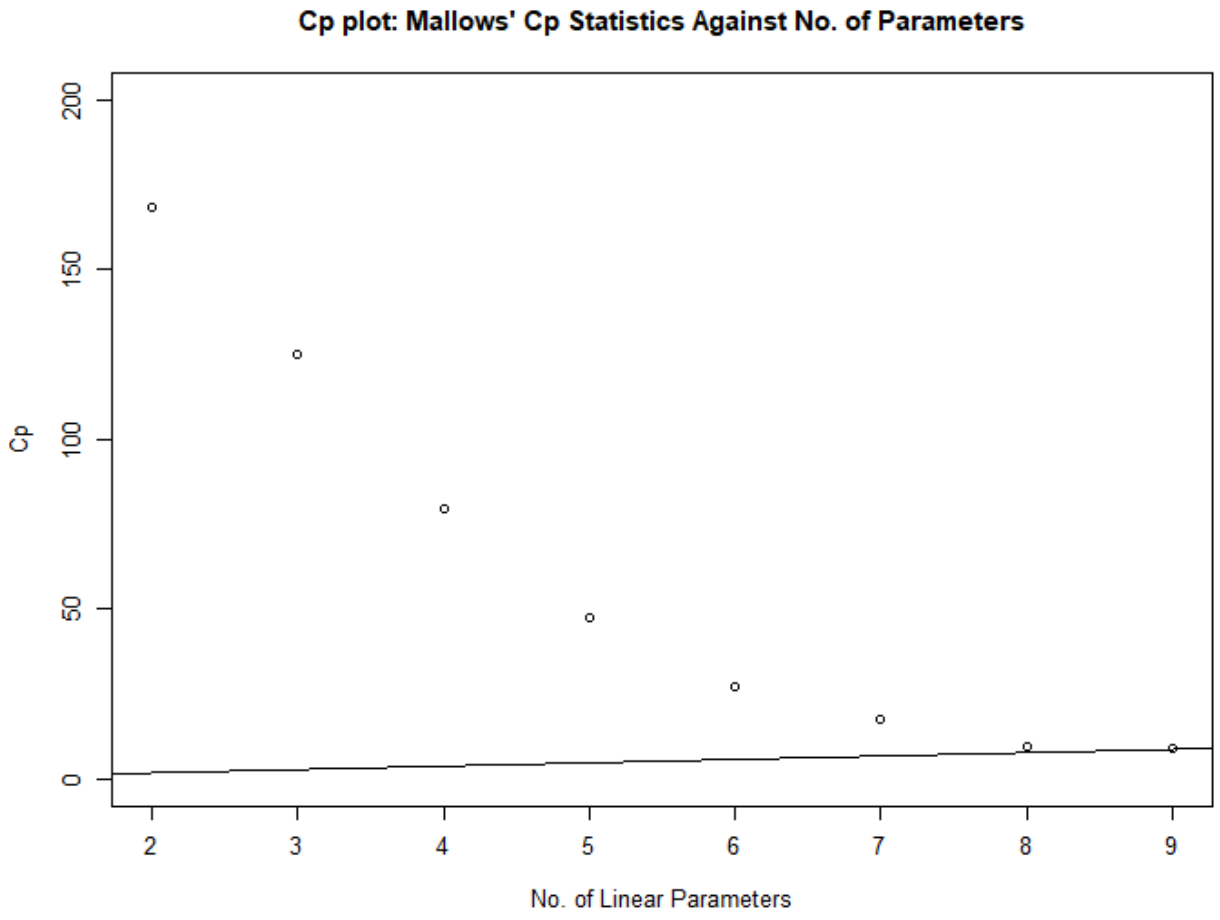

Figure 9: Scatterplot of the Mallows' Cp Statistic with different number of parameters

We prefer the model with p=9 (that is, the full model we carried out) because it has Cp value closest to p. Moreover, it also has the lowest Cp value, therefore having the lowest residual standard error as well.

Therefore, we would reside on the first model we started with, and the regression line to predict house unit price can be expressed as follows:

$$\hat{\text{price}} = -33730 - 0.2876 \times \text{age} + 15.01 \times \text{distanceToStation} + 0.647 \times \text{numberOfStores} + 233.2 \times \text{latitude} + 230 \times \text{longtitude} + 5.023 \times \text{year2013} + 0.3537 \times \text{month} - 0.1236 \times \text{distanceToStation:longtitude}$$

Conclusion

In conclusion, the most fitting model for the data incorporates all variables along with the interaction term (distance to station and longitude). Notably, all examined variables, including transaction date, year, housing age, number of convenience stores within walking distance, distance to the nearest station, latitude, and longitude, exhibit influence on the housing price per unit area in Taipei. However, as evident from the model summary, year, month and the number of convenience stores within walking distance are comparatively less significant than other factors by comparing their p-values. Hence, we posit that transaction date, housing age, distance to the nearest station, latitude, and longitude are the primary determinants influencing house prices. Recognizing these influential factors is crucial for informed decision-making by homebuyers and sellers, enabling them to prioritize features positively contributing to housing prices. Real estate investors can identify profitable opportunities and formulate strategic investment decisions. Policymakers can also leverage these insights to craft effective housing policies and regulations.

References

Kim, K., & Renaud, B. (2009). The global house price boom and its unwinding: An analysis and a commentary. *Housing Studies, 24*(1), 7-24. https://doi.org/10.1080/02673030802550128

Le, T. (2021). *Calculate the distance between locations (with Longitude & Latitude) in Power BI*. Medium. https://medium.com/@trancyle/calculate-the-distance-between-locations-with-longitude-latitude-in-power-bi-d35a18056760